Journal of Next Generation Sequencing & Applications

Araújo et al., Next Generat Sequenc & Applic 2017, 4:2
DOI: 10.4172/2469-9853.1000147

Mini Review OMICS International

Integrating Next Generation Sequencing, Bioinformatics and Cytogenomics in the Study of Brazilian Mammals

Naiara Pereira Araújo, Gustavo Campos Silva Kuhn and Marta Svartman*

Department of General Biology, Institute of Biological Sciences, Presidente Antônio Carlos Avenue, 6627 - Pampulha, 31270-901. Belo Horizonte, Brazil

***Corresponding author:** Marta Svartman, Department of General Biology, Institute of Biological Sciences, Presidente Antônio Carlos Avenue, 6627 - Pampulha, 31270-901. Belo Horizonte, Brazil, Tel: 55(31)34092612; Fax: 55(31)34092567; E-mail: svartmanm@ufmg.br

Rec date: Jun 26, 2017; **Acc date:** July 18, 2017; **Pub date:** July 20, 2017

Abstract

Since the discovery of the DNA double helix, understanding the complexity and diversity of genomes has been a major focus of genetics, including medical and evolutionary research. Sequencing methods have been developed since the 1970's, starting with the first-generation or Sanger sequencing.

Keywords: Next generation sequencing; Bioinformatics; Cytogenomics; Sanger sequencing; DNA-DNA hybridizations

Mini Review

A second-generation or Next-Generation Sequencing (NGS) was developed in 2005 and consisted of a fast speed and high cost-benefit sequencing, when compared to traditional methodologies [1-3]. NGS symbolized a new revolution in genomics, as it has the potential to tackle biological problems related to the genome, transcriptome and epigenome of any species. Albeit the many advantages of this technology, it requires sufficient knowledge of bioinformatics to analyze the data. Moreover, it produces short reads which makes it difficult to assemble whole genomes [4]. Although a third-generation of sequencing that aims at overcoming the limitation of NGS has been evolving since 2008 [5], the second-generation is still largely employed.

Among the diversity of algorithms and bioinformatics tools for the analyses of NGS data, our group has been successfully using RepeatExplorer [6,7], implemented on the galaxy platform, to identify and characterize repetitive DNAs in species with low coverage genome sequencing (Figure 1a).

Once the sequenced data are available, RepeatExplorer can be used to perform all-to-all comparison analyses of sequence reads by similarities and to build clusters of overlapping reads that represent repetitive DNA families (Figure 1b).

This is an easy, fast and more accurate methodology to identify even low abundant repetitive families, sometimes undetected in experimental studies. In the Repeat Explorer output, we look for clusters composed of tandemly repeated satellite DNAs (satDNAs), or transposable elements, complementing the analyses with other softwares such as Tandem Repeats Finder [8] and Dotlet [9], and the databases for repetitive DNAs called Repbase [10]. The bioinformatic analyses are followed by molecular biology experiments, including localization of the repetitive DNAs on chromosome preparations by fluorescent in situ hybridization (FISH; Figure 1b).

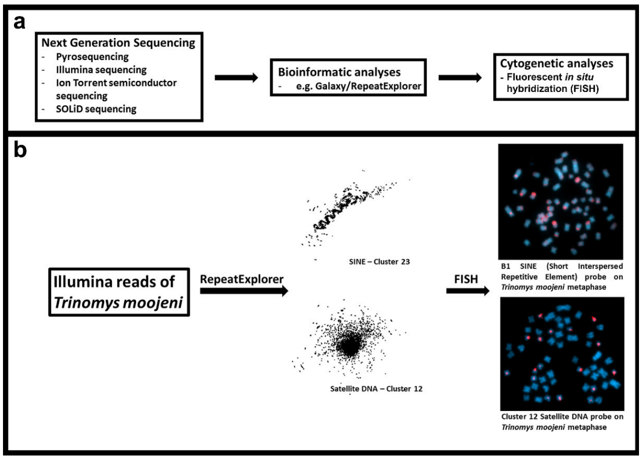
a
Next Generation Sequencing
- Pyrosequencing
- Illumina sequencing
- Ion Torrent semiconductor sequencing
- SOLiD sequencing
Bioinformatic analyses
- e.g. Galaxy/RepeatExplorer
Cytogenetic analyses
- Fluorescent *in situ* hybridization (FISH)
b
Illumina reads of *Trinomys moojeni*
RepeatExplorer
SINE – Cluster 23
Satellite DNA – Cluster 12
FISH
B1 SINE (Short Interspersed Repetitive Element) probe on *Trinomys moojeni* metaphase
Cluster 12 Satellite DNA probe on *Trinomys moojeni* metaphase

Figure 1: (a) An overview of the integration of Next Generation Sequencing technology, bioinformatic and cytogenetic analyses. (b) As an example, we show the steps involved in the analyses of the sequenced genome of Trinomys moojeni (Rodentia: Echimyidae), including the identification of repetitive elements in Repeat Explorer (graph-based clusters) and the localization of two repetitive elements on chromosome preparations by FISH.

We have been using this approach to address several aspects of repetitive DNAs, such as origin, structure, organization, variability, chromosome distribution, and the role of these elements in chromosome/genome evolution and their possible use as taxonomic markers. At the chromosomal level, the distribution of repetitive DNAs may be evolutionarily important, since they have been related to genome and karyotype remodeling in several eukaryotes [11].

The integration of the methodologies described above may be applied to the study of any eukaryote group, and we will use as examples the two mammalian groups in which we have concentrated in the last few years: Brazilian rodents and monkeys. Although both of these groups have been intensively studied, they still present many taxonomic problems, basic cytogenetic information is missing for many taxa, and little is known about the mechanisms related to their genome evolution. We believe that both taxa are especially interesting for genome research, as many of their genera present highly variable karyotypes, including B-chromosomes, multiple sex chromosome systems, and several rearrangements, such as inversions, translocations, fusions/fissions, constitutive heterochromatin variation and centromere repositioning, differentiating even closely related taxa.

For example, we studied Brazilian rodent species of the genus Trinomys (family Echimyidae), typically difficult to identify taxonomically by their morphology and sometimes even by their karyotypes. We investigated the satDNAs of three specimens (Trinomys moojeni, T. setosus, and T. sp.) after sequencing their genome by NGS. A genomic library was prepared using the Nextera Kit, according to the manufacturer's instructions (Illumina Inc., San Diego, CA), and paired-end sequenced in a single flowcell using a MiSeq instrument with the MiSeq Reagent Kit V3 (600 cycles).

Sequencing achieved coverages between 9.1% and 4.5% (T. moojeni), 14.1% and 7% (T. setosus), and 16.7% and 8.3% (T. sp.), assuming 3 pg and 6 pg genome sizes, respectively. Although the sequence data generated had low coverage, it was enough to study the repetitive sequences of these species. Thus, after clustering Illumina reads by similarity, using RepeatExplorer, we identified a satDNA family with 350 bp motifs that showed species-specific features after sequence comparisons and phylogenetic analyses. These species have very similar karyotypes, so we concluded that, although the divergence time among them was not enough to produce karyotypic changes, it was sufficient to allow accumulate some sequence differentiation in this satDNA, which may therefore be used as a species-specific taxonomic marker (Araújo et al. in preparation).

In another work, we recently identified the CarB satDNA in the sequenced genome of the New World monkey Callithrix jacchus [12]. Previous studies based on genomic DNA digestions and DNA-DNA hybridizations led Alves to suggest that this satDNA was specific for Mico species. In fact, we found that CarB represents only 0.1% of the C. jacchus genome and such small amount combined to its sequence divergence might have hindered its detection in the previous experimental studies [13].

Conclusion

In summary, new technologies, like NGS, combined with bioinformatic analyses, have been fostering the identification of a large set of repetitive DNAs that can now be used, for example, in cytogenetic and taxonomic studies. Moreover, the integration of sequenced data with physical mapping on chromosomes of highly diverse groups, like rodents and new world monkeys, promises to greatly contribute to a better understanding of several issues related to chromosome and genome structure and evolution.

References

1. Watson JD, Crick FH (1953) The structure of DNA. Cold Spring Harb Symp Quant Biol 18: 123-131.
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. PNAS 74: 5463-5467.
3. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5: 16-18.
4. Raza K, Ahmad S (2016) Principle, analysis, application and challenges of next-generation sequencing: A review. arXiv preprint arXiv:1606.05254
5. Hayden EC (2009) Genome sequencing: The third generation. Nature 457: 768-769.
6. Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11: 378.
7. Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29: 792-793.
8. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
9. Junier T, Pagni M (2000) Dotlet: Diagonal plots in a web browser. Bioinformatics 16: 178-179.
10. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.
11. Eichler EE, Sankoff D (2003) Structural dynamics of eukaryotic chromosome evolution. Science: 301: 793-797.
12. Araújo NP, De Lima LG, Dias GB, Kuhn GCS, Melo AL, et al. (2017) Identification and characterization of a subtelomeric satellite DNA in Callitrichini monkeys. DNA Res dsx010.
13. Alves G, Canavez F, Seuánez H, Fanning T (1995) Recently amplified satellite DNA in Callithrix argentata (Primates, Platyrrhini). Chromosome Res 3: 207-213.